October 20 Mobile Collaborative Meeting Agenda
Privacy and Open Data Processing Discussion Draft

- Proposed fields to be processed or redacted in technical privacy review
    - *Three Aspects for Focus: Time, Location, Series of characters with identifiable features.*

Past discussions with the collaborative and privacy experts identified three aspects of collected data meriting attention in the technical privacy review. High precision date and time information collected in the execution of the tests provide the opportunity to understand how broadband performance changes in time, but can present privacy concerns. High precision location information derived from the handset's GPS provide the opportunity to understand how broadband performance changes across the nation, but presents potential for concerns that a specific volunteer could be identified with a particular location. In addition to high-precision GPS location information, information about the cellular base station a handset is connected to are also collected and could pose privacy concerns. A variety of character strings are derived from the operating system that can vary subtly in case, punctuation or other features that could present concerns that a specific volunteer could be identified with a particular set of tests.

The program values the privacy of its volunteers and has adopted a conservative approach to mitigating privacy concerns and plans to delete all high precision time and location information for the initial Open Data release. In order to minimize risks while preserving the utility of the Open Data, date and location variables will be processed to preserve temporal and spatial variance among observations in any single batch of tests. Time elapsed, distances, and azimuth between discrete measurements will be calculated and saved as new variables to replace time and location data that is deleted from released data sets. The year and quarter will be calculated to replace the high-precision time of any observation. Not all tests include a high-precision GPS location but valid latitude and longitude values will be replaced with the cellular market area where available. Other information about the source of the GPS observation, the accuracy in meters of the location and other information that poses minimal risk will not be deleted. Cellular base station information, such as the cell tower identifier code that could provide a rough location will be deleted, although other network information may be included.

- Time
    - All data appear with two precise time and date variables (dtime, localdtime). These date and time values will be removed and replaced with four variables
        - Year. The year is extracted from the localdtime and saved as an integer.

- Quarter. The localdtime is processed to identify the date's quarter for the year and saved as an integer.
- Metric Order. In some cases, multiple rows of data share the same submission_id from the same batch of tests, for example where multiple locations or bearer channel changes occur during a test. In order to preserve the ordinality of the rows, the order of all rows of data sharing the same submission_id from the same batch of tests is processed and saved as an integer.
  - row_number() over (partition by submission_id order by localdtime) as metric_order
- Lead Seconds. In order to preserve the amount of time passing between the multiple rows of data sharing the same submission_id from the same batch of tests, the localdtime is converted into seconds and subtracted from the preceding row to determine the number of seconds elapsed between the two rows and saved as an integer.
  - retr_seconds (l.localdtime) - lag (retr_seconds (l.localdtime)) over (partition by l.submission_id order by metric_order) as lead
  - CREATE FUNCTION retr_seconds (timestamp) RETURNS integer AS $$ select ((extract (hour from $1) * 3600)::integer + (extract (minutes from $1) * 60)::integer + (extract (seconds from $1)::integer)) $$ LANGUAGE SQL IMMUTABLE RETURNS NULL ON NULL INPUT;

- Location
  - As tests execute, the operating system can provide the current GPS latitude and longitude, although for a variety of reasons not every test measurement may have a GPS location. Each GPS longitude and latitude will be deleted and replaced with the CMA when available.
    - *Cellular market area ID.* GPS locations can be processed to provide the cellular market area where tests were executed. No additional geographic information will be available for measurements that did not record a valid GPS longitude and latitude.
      - UPDATE curr_location SET geom = ST_SetSRID(ST_MakePoint(longitude,latitude),4326) where extract (year from localdtime) = 2016 and geom is NULL;

        - Update curr_location set cma_id = as c.id from cma_2010 c where curr_location.cma_id is not null and st_within (curr_location.geom, c.geom) ;
    - Cell Tower ID. Base station ID's and Cell Tower IDs may be used to derive a rough location and will be deleted, although other network information may be included.
- Series of characters with identifiable features.
    - Test data that is derived from operating system APIs can expose particular features of a local area network or version of handset or operating system. Because such features could potentially be used to associate test data with a particular handset, data may be processed to remove variations in case or punctuation if variations are associated with low numbers of samples.

- Processed variables
    - Carrier.
        - alter table curr_httpget add column carrier text;
        - update curr_networkdata set carrier = 'att' where Network_operator_code in ('310410') and network_operator_name ~* '^[\s]*at[\s]*[&]*[\s]*t[\s]*$' ;
        - update curr_networkdata set carrier = 'sprint' where network_operator_code IN ('310120') and network_operator_name ~* 'sprint[\s]*$';
        - update curr_networkdata set carrier = 'verizon' where network_operator_code IN ('310012', '311480') and network_operator_name ~* 'verizon';
        - update curr_networkdata set carrier = 'tmobile' where network_operator_code IN ('310026', '310260') and network_operator_name ~* '^[\s]*?t[\-\s]*?mobile';
        - update curr_httpget set carrier= b.carrier from curr_networkdata b where b.metric= 'httpget' and b.carrier is not null and curr_httpget.submission_id = b.submission_id;
        - update curr_httppost set carrier= b.carrier from curr_networkdata b where b.metric= 'httppost' and b.carrier is not null and curr_httppost.submission_id = b.submission_id;
        - update curr_udplatency set carrier= b.carrier from curr_networkdata b where b.metric= 'udplatency' and b.carrier is not null and curr_udplatency.submission_id = b.submission_id;
    - Cellular Test.
        - alter table curr_httpget add column cell boolean not null default FALSE;
        - alter table curr_networkdata add column cell boolean not null default false;

        - update curr_networkdata set cell = 'y' where active_network_type ~* 'mobile' or active_network_type ~* 'cell' or active_network_type ~* 'WIMAX' ;
        - update curr_httpget set cell = 'T' from curr_networkdata b where b.metric= 'httpget' and b.cell = 'T' and curr_httpget.submission_id = b.submission_id;
        - update curr_httppost set cell = 'T' from curr_networkdata b where b.metric= 'httppost' and b.cell = 'T' and curr_httppost.submission_id = b.submission_id;
        - update curr_udplatency set cell = 'T' from curr_networkdata b where b.metric= 'udplatency' and b.cell = 'T' and curr_udplatency.submission_id = b.submission_id;
    - LTE Test.
        - alter table curr_httpget add column lte boolean not null default FALSE;
        - update curr_httpget set lte = 'Y' from curr_networkdata b where b.metric='httpget' and b.network_type ~* 'lte' and curr_httpget.submission_id = b.submission_id ;
    - Temporal and Cross Table Variables.
        - SQL scripts implement the setting of the below variables based on processing of the localdtime or other cross-table joins.

        - alter table curr_httpget add column year integer;
        - alter table curr_httpget add column quarter integer;
        - alter table curr_httpget add column peak boolean not null default FALSE;
        - alter table curr_httpget add column period integer;

        - alter table curr_httpget add column submission_type text;
        - alter table curr_httpget add column carrier text;
        - alter table curr_httpget add column cma_id text;
        - alter table curr_httpget add column cell boolean not null default FALSE;
        - alter table curr_httpget add column lte boolean not null default FALSE;
        - alter table curr_httpget add column roam boolean not null default FALSE;
        - alter table curr_httpget add column android boolean not null default FALSE;